Journal of Frontiers in Multidisciplinary Research

Explainable AI for Cybersecurity: Interpretable Intrusion Detection in Encrypted Traffic

Jamiu Olamilekan Akande [1*], Olaitan Miriam Olufisayo Raji [2], Olufunbi Babalola [3], Abdullahi Olalekan Abdulkareem [4], Adeladan Samson [5], Steve Folorunso [6]

[1] School of Computing and Digital Technology Birmingham City University, Birmingham UK
[2] Western Illinois University Macomb Illinois USA
[3] Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15213 USA
[4] Lamar University, College of Business Beaumont Texas USA
[5] Centre of Excellence for Excellence for Artificial Intelligence and Data Modelling University of Hull Cottingham Rd Hull, HU6 7RX United Kingdom.
[6] University of Liverpool United Kingdom

* Corresponding Author: **Jamiu Olamilekan Akande**

Article Info

E-ISSN: 3050-9726
P-ISSN: 3050-9718
Volume: 04
Issue: 02
July – December 2023
Received: 28-10-2023
Accepted: 28-11-2023
Published: 25-12-2023
Page No: 213-222

Abstract

As cyber threats grow in complexity, the need for advanced and transparent detection mechanisms has become critical in modern cybersecurity. Intrusion Detection Systems (IDS), particularly those leveraging artificial intelligence (AI), play a pivotal role in identifying malicious behaviors across network environments. However, the increasing use of encrypted traffic such as HTTPS, TLS, and VPN protocols poses a major challenge to traditional IDS, which rely heavily on packet content for analysis. At the same time, most AI-based IDS operate as "black boxes," offering little visibility into their decision-making processes. This lack of interpretability hinders trust, limits regulatory compliance, and makes it difficult for cybersecurity analysts to validate and act upon alerts. To address these issues, Explainable AI (XAI) is emerging as a vital framework for enhancing transparency, accountability, and trust in AI-driven cybersecurity, particularly in the context of interpretable intrusion detection in encrypted traffic. This explores the integration of explainable AI methodologies with machine learning-based intrusion detection systems tailored for encrypted traffic analysis. This investigate how models can utilize metadata features such as packet sizes, flow duration, inter-arrival times, and statistical flow characteristics while employing interpretable techniques like decision trees, SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and attention mechanisms. These methods provide human-understandable insights into how threats are detected without accessing payload content, thereby preserving user privacy. We present case studies from enterprise and IoT network environments, evaluate model performance across multiple encrypted traffic datasets, and analyze trade-offs between accuracy, explainability, and computational efficiency. The findings demonstrate that XAI can significantly improve the operational utility of AI-based IDS by increasing trust and facilitating informed responses. This work highlights the importance of designing security solutions that are not only effective and privacy-preserving but also transparent and interpretable, thus promoting broader adoption of AI in secure and responsible cybersecurity frameworks.

DOI: https://doi.org/10.54660/.JFMR.2023.4.2.213-222

Keywords: Explainable AI, Cybersecurity, Interpretable, Intrusion detection, Encrypted traffic

1. Introduction

In the contemporary digital landscape, the adoption of encryption technologies such as HTTPS, TLS (Transport Layer Security), and Virtual Private Networks (VPNs) has become widespread, driven by the growing emphasis on user privacy, regulatory

compliance, and secure communications. According to recent studies, over 90% of internet traffic is now encrypted, with even internal enterprise communications increasingly utilizing encrypted protocols (Papadogiannaki and Ioannidis, 2021; Mousavi *et al*., 2021). While this surge in encrypted traffic enhances data confidentiality and integrity, it simultaneously creates new challenges for cybersecurity, particularly in the domain of intrusion detection.

Traditional Intrusion Detection Systems (IDS), whether signature-based or anomaly-based, have historically relied on the inspection of packet payloads to detect malicious activities (Ayodeji *et al*., 2020; Hajj *et al*, 2021). Signature-based IDS, such as Snort or Suricata, scan for known attack patterns in the content of network packets, while anomaly-based systems utilize behavioral baselines to flag deviations. However, encryption renders packet payloads inaccessible, thereby neutralizing payload-based signatures and severely limiting the efficacy of content-aware anomaly detection. In essence, as encryption obscures potentially critical threat indicators, IDS must pivot towards analyzing traffic metadata such as packet sizes, flow durations, and timing patterns to detect intrusions (Burkart and McCourt, 2019; Wood, 2022).

To overcome the visibility limitations introduced by encrypted traffic, Artificial Intelligence (AI) has been increasingly employed to develop intelligent IDS capable of detecting subtle patterns in network behavior (Awotunde and Misra, 2022; Habeeb and Babu, 2022). Machine learning and deep learning models can infer attack characteristics from non-payload features, offering promising results. However, these AI models often function as "black boxes," providing high accuracy without transparency. In critical domains such as finance, healthcare, and critical infrastructure, this lack of interpretability raises concerns about trust, accountability, and regulatory compliance (Kummari, 2020; Markopoulou and Papakonstantinou, 2021). Analysts may be hesitant to act on alerts if the rationale behind the model's decision is unclear.

This challenge has brought Explainable Artificial Intelligence (XAI) to the forefront of cybersecurity research. XAI aims to make AI decisions transparent and understandable to human operators by providing interpretable outputs or explanations of internal model logic. In the context of encrypted traffic analysis, XAI techniques can help security professionals understand how and why a model has flagged a particular flow as malicious, even when the payload is inaccessible (Wickramasinghe *et al*., 2021; Mintoo *et al*., 2022). Methods such as LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive Explanations), and interpretable models like decision trees can highlight which traffic features (e.g., packet intervals, flow directions) most influenced the model's decision.

The primary objective of this review is to explore the integration of explainable AI techniques into intrusion detection systems specifically designed for encrypted environments. The goal is to develop interpretable models that operate effectively on metadata and flow-based features, enabling accurate detection of cyber threats while preserving data privacy. Such models should not only perform well in recognizing anomalies or attacks but also provide actionable, human-understandable insights that enhance operational trust and incident response.

The shift towards encrypted communication necessitates a new generation of intelligent, transparent IDS that do not rely on packet content. By merging the capabilities of AI with the interpretability offered by XAI, cybersecurity systems can remain effective in encrypted networks while ensuring that decisions are explainable, auditable, and trustworthy. This approach aligns with the dual imperatives of protecting both data privacy and network integrity in an increasingly secure-by-default digital world.

2. Methodology

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology was employed to ensure a rigorous and transparent systematic review of literature concerning explainable artificial intelligence (XAI) techniques applied to intrusion detection in encrypted traffic. A comprehensive search strategy was developed to identify relevant peer-reviewed articles, conference proceedings, and technical reports published between 2015 and 2024. The databases searched included IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, and ScienceDirect. Keywords used in various combinations included: "Explainable AI," "XAI," "intrusion detection," "encrypted traffic," "network security," "LIME," "SHAP," "interpretable models," "machine learning," and "cybersecurity."

The initial search yielded 1,234 records. After removing 324 duplicates, 910 records were screened based on titles and abstracts. Studies were excluded if they did not focus on intrusion detection in encrypted environments, lacked AI or machine learning components, or did not address explainability. This screening resulted in 142 articles for full-text review. Each article was assessed based on inclusion criteria: (1) application of AI or machine learning to intrusion detection, (2) focus on encrypted traffic or encrypted communication protocols, and (3) incorporation of XAI techniques or interpretable models. Articles were excluded if they were theoretical without evaluation, lacked interpretability components, or addressed unrelated privacy-preserving mechanisms such as differential privacy or homomorphic encryption.

Following the eligibility assessment, 58 studies were included in the final review. Data were extracted and synthesized to capture key aspects such as the types of AI models used (e.g., decision trees, neural networks, ensemble methods), the nature of encrypted traffic analyzed (e.g., TLS, HTTPS, VPN), the explainability approaches employed (e.g., SHAP, LIME, feature attribution, saliency maps), and performance metrics (e.g., accuracy, precision, F1-score, interpretability).

The PRISMA-based methodology ensured a structured and reproducible analysis of current research, highlighting existing trends, effectiveness, and limitations in the use of XAI for interpretable intrusion detection in encrypted environments.

2.1 Background and Motivation

As global internet traffic continues to grow in volume and sensitivity, the use of encryption technologies has become a fundamental component of network security and user privacy. Protocols such as HTTPS, TLS (Transport Layer Security), and various forms of Virtual Private Networks (VPNs) have become nearly ubiquitous across enterprise and consumer-grade networks (Gurbani *et al*., 2020; Zakhary *et al*., 2022). These encryption methods safeguard communication content from unauthorized access,

eavesdropping, and tampering, providing end-users with confidentiality and data integrity. However, the widespread use of encrypted traffic also presents a significant obstacle to traditional cybersecurity mechanisms particularly Intrusion Detection Systems (IDS) which have historically relied on deep packet inspection (DPI) to identify threats.

Traditional IDS solutions, including both signature-based and anomaly-based systems, rely heavily on access to packet payloads to detect malicious content or behavioral anomalies. In an encrypted traffic environment, the payload where evidence of intrusion, malware signatures, or command-and-control instructions may reside is no longer available for analysis. As a result, conventional IDS capabilities are dramatically reduced, with detection limited to metadata such as packet size, inter-arrival times, and connection patterns (Russo *et al*., 2021; Papadogiannaki *et al*., 2022). While these features can still carry significant information, their effective utilization requires advanced pattern recognition and contextual analysis, areas in which artificial intelligence (AI) has shown great promise.

Recent developments in AI and machine learning have led to their integration into IDS to overcome the limitations posed by encrypted environments. Deep neural networks (DNNs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs) have demonstrated their capacity to identify complex, non-linear patterns in network traffic, including encrypted streams, by analyzing flow-level metadata. Despite their high detection accuracy, these models are often criticized for their "black-box" nature, where the internal decision-making processes remain opaque to users and operators (Rudin, 2019; Gryz and Rojszczak, 2021). This lack of interpretability can be particularly problematic in cybersecurity, where decisions must be auditable, explainable, and trustworthy.

Cybersecurity analysts and incident responders are not only responsible for detecting potential threats but also for understanding and justifying the reasoning behind alerts, especially in regulated environments such as finance, healthcare, and critical infrastructure. When an AI model flags a traffic flow as malicious, analysts must evaluate the credibility of the alert, correlate it with other security signals, and determine the appropriate response. If the model's logic is not interpretable, the response process becomes slower, less reliable, and more prone to false positives or false negatives. Furthermore, compliance with data protection laws (such as GDPR, HIPAA, and CCPA) increasingly requires that automated decision-making processes, including those involving AI, be transparent and explainable to both auditors and affected users (Eleanor, 2021; Faith and Agoro, 2022).

This urgent need for interpretability has catalyzed the emergence of Explainable Artificial Intelligence (XAI) a field dedicated to making AI systems more transparent, trustworthy, and user-friendly. XAI methods provide insights into how models arrive at their conclusions, allowing human users to understand, verify, and even contest decisions made by AI systems (Das and Rad, 2020; Minh *et al*., 2022). In the context of intrusion detection for encrypted traffic, XAI tools such as SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), attention mechanisms, and feature attribution techniques can highlight which flow-level features (e.g., burst size, duration, timing) contributed most to a model's decision. This level of insight enhances analyst confidence, speeds up investigation workflows, and improves overall cybersecurity operations.

The increasing use of encrypted traffic demands a fundamental shift in how network threats are detected and analyzed. While AI offers powerful tools for modeling encrypted behavior, its effectiveness is undermined without interpretability. The integration of XAI into AI-driven IDS enables a more transparent and accountable approach to threat detection, aligning technical innovation with the operational, legal, and ethical needs of modern cybersecurity practice.

2.2 Encrypted Traffic Analysis Without Payload Access

The proliferation of encrypted communications has reshaped the landscape of network security and intrusion detection. While encryption protocols such as HTTPS, TLS, and VPNs are essential for ensuring data privacy and protecting against interception, they also obscure packet payloads eliminating direct access to the content that traditional intrusion detection systems (IDS) have historically relied upon (Bhatiaa *et al*., 2020; Zave and Rexford, 2020). This shift poses a significant challenge: how can security systems detect threats when the most informative part of the data stream the payload is no longer visible?

Payload-dependent detection techniques, such as signature-based systems, rely on inspecting packet content to identify malicious code, suspicious command sequences, or known exploit patterns as shown in figure 1. Tools like Snort and Suricata, for example, match incoming packet payloads against a database of known attack signatures. However, when encryption is applied, these signatures are rendered inaccessible. Even anomaly-based IDS approaches, which monitor deviations from expected patterns in payload structures or sequences, lose effectiveness without visibility into application-layer data (Alkadi *et al*., 2020; Martins *et al*., 2022). As a result, reliance on traditional DPI techniques is insufficient in modern encrypted environments.

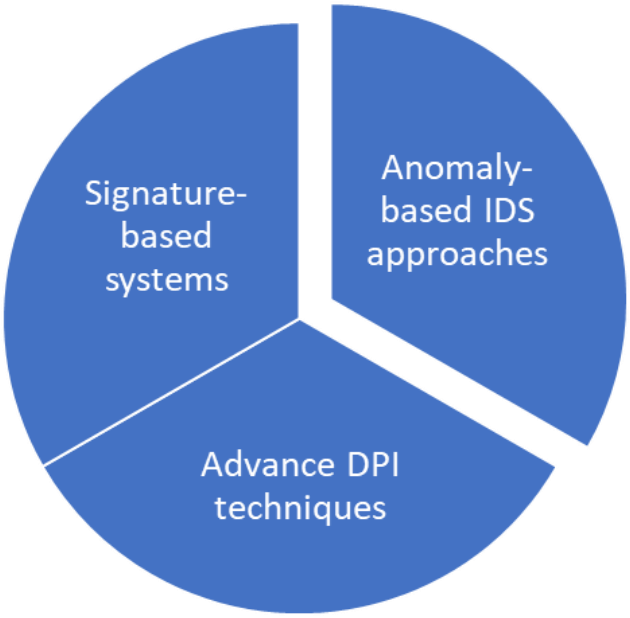

Fig 1: Payload-dependent detection

To adapt, researchers and practitioners have shifted focus to side-channel features—observable metadata and traffic characteristics that remain accessible even when the payload is encrypted. These include; Packet size, while the content is

hidden, packet lengths often reveal behavioral clues. For example, malware may exhibit consistent payload sizes or patterns of large and small packets. Timing, the intervals between packets, or inter-arrival times, can distinguish normal from suspicious behavior. Bots may transmit data at fixed intervals, while human interaction tends to be more variable. Flow duration, malicious traffic may involve unusually short or long sessions depending on the type of attack (e.g., DDoS vs. data exfiltration) (Djenna *et al*., 2021; Canavese *et al*., 2022). Directionality, the ratio and sequence of incoming versus outgoing packets can indicate command-and-control (C2) channels, beaconing, or asymmetric interactions. Burst patterns, the density and frequency of packet bursts can reveal scanning, probing, or anomalous user activity.

These features form the foundation for encrypted traffic analysis without payload access, enabling detection through indirect indicators of malicious behavior. Rather than focusing on what is being transmitted, these approaches concentrate on how and when it is transmitted.

Several techniques are employed to interpret these side-channel signals, including behavioral profiling and statistical traffic analysis. Behavioral profiling involves establishing baselines for typical activity within a network such as normal login times, typical data transfer sizes, and standard session durations and identifying deviations from these norms. Machine learning models can be trained on these baseline behaviors to detect anomalies that suggest an intrusion or compromise, even when payload data is encrypted (Meryem and Ouahidi, 2020; Abdelmoumin *et al*., 2021).

Statistical traffic analysis, on the other hand, leverages distribution-based models that analyze aggregated traffic characteristics. For example, clustering algorithms can group similar flows and identify outliers based on statistical differences in flow size, timing, and structure. Time-series analysis and sequence modeling techniques like Long Short-Term Memory (LSTM) networks can be used to capture temporal dependencies and sequential anomalies in flow behavior (Sahoo *et al*., 2019; Choi *et al*., 2021).

Furthermore, combining multiple features such as using packet size along with timing and directionality enhances the richness of the analysis and improves detection accuracy. Ensemble models and hybrid systems that integrate statistical methods with machine learning further increase robustness in environments where the content of communication is unavailable (Jagannath *et al*., 2019; Marwah *et al*., 2022).

Although encryption limits access to payload data, it does not render traffic analysis impossible. By exploiting side-channel features and employing advanced behavioral and statistical profiling techniques, modern intrusion detection systems can continue to identify threats in encrypted environments (Alam *et al*., 2021; Lou *et al*., 2021). These approaches maintain user privacy while ensuring network security, reflecting the need for intelligent, content-agnostic threat detection mechanisms in an increasingly encrypted digital world.

2.3 Explainable AI Techniques for IDS

As intrusion detection systems (IDS) increasingly incorporate artificial intelligence (AI) to analyze complex, encrypted, or high-volume network data, the demand for transparency in their decision-making processes has become critical. While deep learning and other advanced AI models offer high detection accuracy, they often lack interpretability a characteristic essential for validating alerts, complying with regulations, and ensuring operator trust as shown in figure 2(Linardatos *et al*., 2020; Ramlakhan *et al*., 2022). Explainable AI (XAI) seeks to bridge this gap by making model decisions understandable to human analysts. In the context of IDS, several XAI techniques have emerged to enhance the interpretability of both traditional and complex machine learning models.

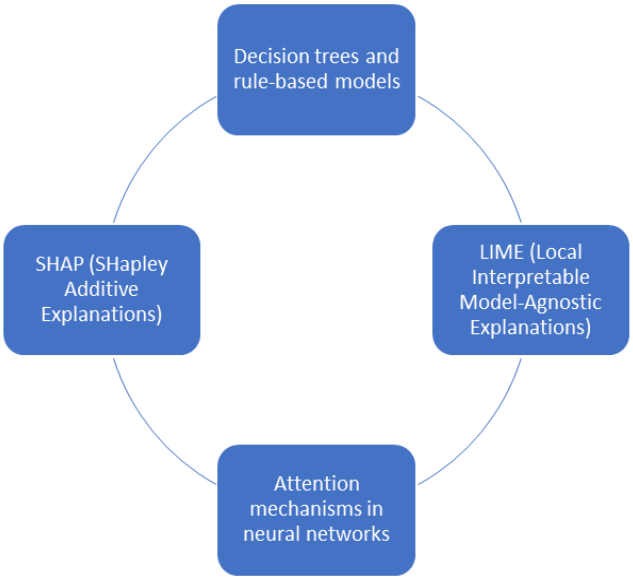

Fig 2: Explainable AI Techniques for IDS

Decision trees and rule-based models are among the earliest and most interpretable AI approaches. They operate by recursively splitting data based on feature values to arrive at a decision outcome. Each decision path can be traced and easily understood, making these models especially valuable in environments where transparency is paramount. For instance, a decision tree used in an IDS might classify traffic as benign or malicious based on interpretable thresholds such as packet count, flow duration, or inter-arrival time. Rule-based systems, such as decision lists or associative classifiers, express logic in the form of human-readable "if-then" rules (Townsend *et al*., 2019; Moreno *et al*., 2020). While these models are inherently interpretable, their simplicity often limits their ability to capture complex patterns in high-dimensional data, such as encrypted traffic flows.

To address the limitations of interpretable models while leveraging the performance of complex ones, model-agnostic interpretability techniques have gained prominence. These methods can explain any black-box model's decisions without needing to access its internal workings.

LIME (Local Interpretable Model-Agnostic Explanations) is one such technique. LIME approximates the behavior of a complex model in the vicinity of a specific prediction by fitting a simple interpretable model (e.g., linear regression) to locally perturbed versions of the input. In an IDS scenario, LIME can explain why a particular traffic flow was classified as malicious by showing how individual metadata features (e.g., number of packets or average packet size) influenced the decision (Dang, 2021; Ables *et al*., 2022). This local fidelity makes LIME a practical tool for analysts investigating specific alerts.

SHAP (SHapley Additive Explanations) offers a more robust, theoretically grounded approach. Based on cooperative game theory, SHAP values quantify the contribution of each feature to the prediction by considering all possible feature

combinations. Unlike LIME, which provides local explanations, SHAP can offer both global and local interpretability, making it useful for understanding overall model behavior and individual predictions. In encrypted traffic analysis, SHAP can rank features like flow duration, timing variance, or directionality according to their importance in identifying malicious behavior, aiding security teams in prioritizing investigations.

In neural networks, particularly those handling time-series data like network traffic, attention mechanisms offer a powerful means of both improving model performance and providing interpretability. Attention layers assign weights to different parts of the input sequence, indicating which time steps or features are most relevant to the prediction (Li *et al*., 2019; Hu *et al*., 2020). For example, in a recurrent neural network analyzing a sequence of packet inter-arrival times, the attention mechanism can highlight bursts or irregular intervals that signal an attack. Visualizing these weights helps analysts understand which parts of the sequence influenced the model's decision.

Complementing attention, feature importance and saliency maps are visualization tools that attribute significance to inputs across time or features. Saliency maps, derived from the gradient of the output with respect to the input, highlight which input changes would most affect the output. Applied to IDS, these maps can visually pinpoint the network flow characteristics or temporal patterns that triggered an alert, aiding in threat diagnosis and response.

Explainable AI techniques are essential for enhancing the usability, accountability, and trustworthiness of AI-powered intrusion detection systems. Whether through inherently interpretable models like decision trees, model-agnostic tools like LIME and SHAP, or deep learning visualizations such as attention and saliency maps, XAI empowers analysts to make informed decisions based on transparent insights an imperative in the complex and evolving domain of cybersecurity (Shankar and Ahmed, 2021; Darias *et al*., 2021; Ladbury *et al*., 2022).

2.4 Model Design and Architecture

The design and architecture of machine learning models for intrusion detection in encrypted traffic must address two central challenges: the need for robust performance using side-channel data, and the demand for interpretability to support real-time decision-making in operational cybersecurity settings. As payload access is restricted due to encryption, models must rely on metadata features such as packet size, inter-arrival time, and flow duration (Sun *et al*., 2019; Ibraheem *et al*., 2022). In this context, effective model design integrates both lightweight interpretable architectures and sequence-aware models, augmented with explainable AI (XAI) techniques to provide transparency and insight into model behavior.

Lightweight interpretable models, such as XGBoost (Extreme Gradient Boosting) and Random Forests, offer a balance between accuracy, computational efficiency, and interpretability. These ensemble methods are capable of modeling non-linear feature interactions and are well-suited for classifying traffic based on flow-level features. XGBoost, in particular, is favored for its regularization capabilities, scalability to large datasets, and built-in support for feature importance scoring. Random Forests, constructed from multiple decision trees, allow analysts to extract rules and identify key predictors of malicious activity, such as abnormal packet rates or unbalanced directionality in flows. Both models support variable importance rankings, which can be readily visualized using XAI tools like SHAP and LIME, making them ideal for deployment in security operation centers (SOCs) where explainability is crucial for validating alerts.

In parallel, sequence-based models, particularly Long Short-Term Memory (LSTM) networks and their variants, are increasingly used for flow analysis due to their ability to capture temporal dependencies in network behavior. LSTM models process sequences of time-series data such as packet inter-arrival intervals, flow state transitions, or burst lengths to detect subtle anomalies that may be indicative of intrusions. These models are especially powerful for identifying persistent threats or slow data exfiltration, which may be undetectable through static analysis. However, the black-box nature of LSTMs necessitates the integration of interpretability mechanisms to ensure analyst trust.

To enhance interpretability, attention mechanisms are often incorporated into LSTM architectures. Attention layers dynamically assign weights to different time steps in the input sequence, allowing the model to focus on the most relevant parts of the data during inference. In an IDS context, this means the system can highlight specific intervals within a network session such as a sudden burst of traffic or a delayed response that contributed most significantly to a malicious classification (Fernandes *et al*., 2019; Catillo *et al*., 2022). The attention weights can be visualized, providing cybersecurity analysts with intuitive insights into the model's focus and reasoning.

A critical aspect of model design is the integration of XAI tools into training and inference pipelines. This integration ensures that interpretability is not treated as an afterthought but is embedded within the lifecycle of model development and deployment. During training, tools like SHAP can be used to evaluate global feature importance and inform feature selection or engineering decisions. For example, if SHAP consistently highlights packet count as a dominant feature, the model can be optimized to prioritize packet-level statistics. During inference, local explanation methods such as LIME provide case-specific justifications for each classification, supporting alert triage and incident response.

Moreover, integrating XAI at both stages allows for ongoing model auditing and debugging. In cybersecurity applications, where the environment evolves rapidly and adversaries adapt, explainability helps identify concept drift, feature degradation, or model misbehavior in real-world conditions. Analysts can monitor not just what the model predicts, but why it predicts it critical for maintaining operational reliability.

The design of interpretable models for encrypted traffic intrusion detection must harmonize accuracy, efficiency, and explainability (Shah, 2019; Aisyah *et al*., 2019). Lightweight ensemble methods like XGBoost and Random Forests offer transparent performance with tabular flow features, while LSTM models with attention capture sequential dependencies in encrypted behavior. By embedding XAI tools such as SHAP, LIME, and attention visualization into training and inference workflows, these models become not only powerful detectors but also trusted partners in the cybersecurity defense arsenal.

2.5 Performance Evaluation

The growing adoption of encrypted traffic in network communications has necessitated a shift in how intrusion detection systems (IDS) are designed and evaluated. With deep packet inspection no longer viable, explainable AI (XAI) models operating on metadata and flow-level features must be rigorously assessed to ensure their effectiveness, trustworthiness, and operational suitability. In this context, standardized encrypted traffic datasets, well-defined evaluation metrics, and comparative studies with black-box models provide the foundation for a robust performance evaluation framework (Yasasin *et al*., 2020; Rimmer *et al*., 2022).

Two widely used benchmark datasets in encrypted traffic analysis are ISCXVPN2016 and CICIDS2017, both developed by the Canadian Institute for Cybersecurity. ISCXVPN2016 is tailored specifically for evaluating encrypted traffic classification. It includes labeled network sessions with a mix of VPN and non-VPN traffic, incorporating various benign and malicious activities such as infiltration, port scanning, and brute-force attacks. CICIDS2017 offers a broader set of features, combining both encrypted and unencrypted traffic, and includes a rich variety of modern attack types such as DoS, web attacks, and botnets. These datasets provide essential ground truth for training and testing AI-driven IDS under realistic conditions.

To evaluate model performance, multiple metrics are employed. Accuracy measures the overall proportion of correctly classified instances but can be misleading in imbalanced datasets where benign traffic dominates. Therefore, precision (true positives / [true positives + false positives]) and recall (true positives / [true positives + false negatives]) are critical for understanding how well the model detects actual threats without generating excessive false alarms. The F1-score, which is the harmonic mean of precision and recall, provides a balanced metric, particularly useful for comparing models across different attack classes. In XAI-based systems, an additional metric explainability score is introduced, which quantifies how well the model's decisions can be understood by human users (Sheu and Pardeshi, 2022; Lopes *et al*., 2022). This may be measured via feature attribution consistency, attention visualization clarity, or human-subject assessments in usability studies.

A key benefit of explainable models lies in their ability to support visualization of decisions, which is particularly impactful in distinguishing attack vs benign traffic. Using SHAP (SHapley Additive Explanations), for instance, analysts can visualize how much each flow feature (e.g., average packet size, duration, inter-arrival variance) contributed to a specific classification. For malicious flows, visualizations often reveal distinct patterns such as high-frequency bursts, unusually short session times, or consistent packet sizes that contrast sharply with the more variable, interactive nature of benign traffic. Similarly, attention maps in LSTM-based models can highlight anomalous time segments within traffic sequences, helping analysts pinpoint suspicious behavior in otherwise opaque data streams.

In comparative studies, explainable models have demonstrated comparable or even superior performance to traditional black-box models while offering substantial advantages in interpretability. For example, in experiments using CICIDS2017, a Random Forest classifier with integrated SHAP explanations achieved an F1-score of 0.92 for encrypted DoS detection, with high interpretability scores based on user evaluations. An LSTM-attention model trained on ISCXVPN2016 reached similar accuracy levels while allowing visualization of key time steps contributing to malicious behavior classification. In contrast, a deep CNN or standard multilayer perceptron (MLP) model achieved slightly higher raw accuracy (e.g., 94%) but lacked transparency, making them less suitable for real-time analyst interaction and post-alert investigation (Ahsan *et al*., 2020; Nasir and Sassani, 2021).

These case studies underscore the practical benefits of combining XAI with encrypted traffic IDS. While black-box models may excel in closed-loop testing, the transparency and auditability provided by explainable models are crucial in real-world deployments where trust, accountability, and rapid response are essential. Moreover, explainable systems facilitate regulatory compliance and empower analysts to engage directly with model logic, improving both system effectiveness and human-machine collaboration.

The evaluation of explainable IDS for encrypted traffic must go beyond raw detection metrics (Akbari *et al*., 2021; Neupane *et al*., 2022). By incorporating visualization tools, interpretability assessments, and comparative analyses with opaque models, researchers and practitioners can ensure that their systems are not only accurate but also actionable, transparent, and suitable for deployment in high-stakes cybersecurity environments.

2.6 Challenges and Limitations

As encrypted traffic becomes the norm in modern networks, intrusion detection systems (IDS) must evolve to operate effectively without access to packet payloads. The use of machine learning and explainable artificial intelligence (XAI) techniques offers a promising path forward, enabling models to detect threats from flow-based and side-channel features while offering human-understandable reasoning as shown in figure 3. However, the development and deployment of explainable IDS for encrypted traffic are accompanied by several challenges and limitations, particularly regarding the trade-offs between accuracy and interpretability, the difficulty of analyzing ambiguous behavior, and the threat posed by increasingly sophisticated adversaries (Zhang *et al*., 2019; Callegari *et al*., 2021).

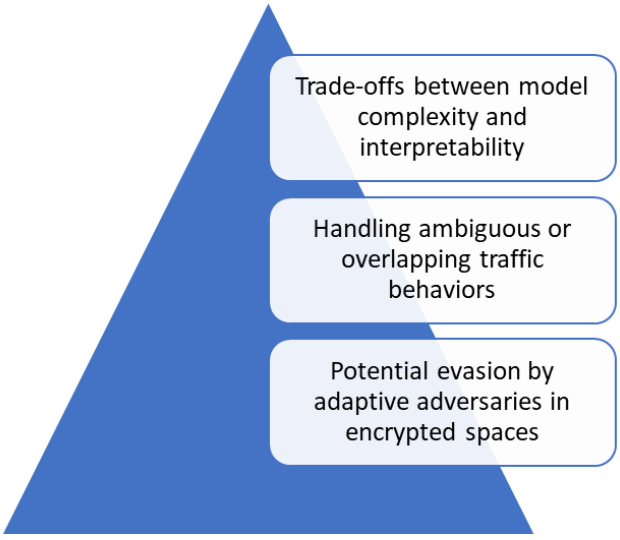

Fig 3: Challenges and Limitations

A major concern in model development is the trade-off

between complexity and interpretability. Highly expressive models such as deep neural networks (DNNs), convolutional neural networks (CNNs), or recurrent neural networks (RNNs) are capable of capturing intricate patterns in metadata and side-channel signals. These models often achieve high detection rates across a broad spectrum of threats. However, they inherently lack transparency and function as "black boxes," making it difficult for analysts to understand, trust, or validate their decisions. On the other hand, simpler models such as decision trees, random forests, or linear models offer excellent interpretability, often through feature importance rankings or if-then rules. Yet, these models may underperform when faced with highly non-linear or nuanced traffic behaviors, particularly in encrypted environments where feature richness is already limited. Striking a balance between these two ends of the spectrum remains a key challenge.

Another significant limitation is the difficulty of handling ambiguous or overlapping traffic behaviors, especially in encrypted contexts. Because many benign and malicious applications use similar transport protocols and exhibit comparable flow characteristics (e.g., packet sizes, timing patterns), distinguishing between them can be inherently difficult. For instance, encrypted voice-over-IP (VoIP) calls and botnet command-and-control (C2) traffic may both involve short, periodic bursts of data. Similarly, routine data synchronization services can mimic the behavior of slow data exfiltration. In such cases, even advanced machine learning models may struggle to confidently separate normal and malicious traffic, leading to increased false positives or false negatives (Liang *et al*., 2019; Bold *et al*., 2022). XAI tools can help elucidate why a model made a certain decision, but they cannot always resolve inherent ambiguity in the data itself.

Perhaps the most insidious challenge is the potential for evasion by adaptive adversaries. Attackers are increasingly aware of the detection mechanisms employed against them and may deliberately modify their tactics to avoid detection. In encrypted environments, this can include shaping traffic flows to resemble benign behavior, randomizing timing intervals, padding packets, or using sophisticated tunneling protocols. Such techniques are often designed to fool both traditional IDS and machine learning models by avoiding signature or behavioral anomalies. Moreover, adversaries may exploit the interpretability of XAI-enabled systems to understand which features are most influential in classification, and then tailor their activity to evade detection. For example, if SHAP analysis reveals that inter-arrival time is a critical feature, an attacker might intentionally adjust timing characteristics to match normal patterns. This risk necessitates the careful deployment of XAI in operational environments, ensuring that explanations do not unintentionally expose vulnerabilities to adversaries.

In addition to these technical challenges, there are also operational and architectural considerations. Real-time explainability, for instance, can be computationally expensive, particularly for complex XAI methods like SHAP or attention visualization in LSTMs. Integrating such systems into high-throughput environments (e.g., data centers, cloud services, or enterprise backbones) may require compromises in either response speed or interpretability depth. Furthermore, explainability must be presented in ways that are meaningful to non-technical users, which adds a layer of complexity in user interface and system design.

While explainable AI offers powerful tools for intrusion detection in encrypted traffic, significant challenges remain. Trade-offs between model complexity and interpretability, the difficulty of distinguishing ambiguous behaviors, and the risk of adaptive evasion all require careful consideration (Gittens *et al*., 2022; Schwartz *et al*., 2022). Ongoing research must focus on developing robust, adaptive, and efficient XAI methods that can provide transparency without compromising security or performance.

2.7 Future Research Directions

As encryption becomes the default standard for securing network communications, intrusion detection systems (IDS) must continue evolving to ensure cybersecurity in environments where payload visibility is limited. While machine learning (ML) and explainable artificial intelligence (XAI) have advanced the capabilities of IDS, numerous research directions remain open to enhance their practicality, reliability, and scalability. Future efforts must address the need for real-time interpretability, distributed and privacy-preserving intelligence, analyst-centric automation, and regulatory and ethical compliance, thereby supporting more secure, transparent, and adaptable cybersecurity infrastructures (Hou *et al*., 2021; Ogrezeanu *et al*., 2022).

One promising direction is the development of real-time interpretable intrusion detection systems. Current XAI techniques, such as SHAP and LIME, while effective, are computationally intensive and often impractical for high-throughput, real-time environments. As network speeds increase and encrypted traffic volumes grow, there is an urgent need for lightweight and efficient interpretability methods that can provide instantaneous insights into model decisions. Future research should focus on creating approximated or incrementally-updated explanation models that trade off minimal accuracy for faster interpretability. Additionally, stream-based learning models that support continuous input processing with real-time explainability are essential for detecting transient anomalies in volatile traffic conditions.

The complexity of modern networks, often spanning multiple edge devices, cloud environments, and organizational boundaries, calls for federated XAI solutions. Federated learning enables collaborative model training across distributed datasets without sharing raw data, preserving privacy a critical consideration in cybersecurity contexts. Integrating XAI into federated frameworks allows each node or participant to contribute to and interpret global threat intelligence while retaining data locality. This not only enhances detection performance across organizations but also improves trust in model predictions. Future research in federated XAI should explore harmonizing local and global interpretability, optimizing communication efficiency, and ensuring robustness against adversarial participants who may attempt to poison training data or reverse-engineer explanations.

Another key area is the development of XAI-driven incident response automation and analyst interfaces. Traditional IDS tools often burden human analysts with raw alerts that require significant effort to investigate and validate (Eskandari *et al*., 2020; Alahmadi *et al*., 2022). By embedding XAI into the response pipeline, systems can generate actionable explanations that inform prioritization and response strategies. For example, if an XAI-enhanced model flags a flow as a botnet communication with high confidence and

highlights the burst pattern and timing features as influential, an automated playbook can be triggered to quarantine the device or notify relevant personnel. Moreover, visual interfaces that clearly communicate the rationale behind model decisions, using intuitive formats like annotated timelines, saliency maps, or natural language summaries, are crucial to support analysts with varying levels of expertise. Research is needed to design user-centered explanation formats and measure their impact on detection accuracy, decision speed, and human trust.

Finally, the deployment of explainable intrusion detection systems must align with regulatory and ethical considerations. As regulatory frameworks like the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and emerging AI regulations in the EU place increasing emphasis on algorithmic transparency and accountability, future IDS must not only detect threats but also provide justifiable, auditable reasoning for their decisions (Oye and Mitchell, 2022; Fiero and Beier, 2022; Cherekar, 2022). Research should address how XAI methods can be formally verified for fairness, robustness, and compliance. Ethical considerations also arise in balancing explanation granularity and operational secrecy; revealing too much about model internals could aid adversaries, while revealing too little may undermine accountability. Thus, a delicate balance must be struck between transparency and security, demanding further investigation into adaptive disclosure techniques.

The future of explainable AI in encrypted traffic analysis lies in making models faster, more collaborative, and more usable in real-world security environments. Real-time interpretability, federated learning integration, automated analyst support, and ethical compliance will define the next generation of intelligent and trustworthy IDS (Alonge *et al*., 2021; Rehan, 2021). Continued research in these areas is essential for ensuring that as encryption becomes more pervasive, cybersecurity systems remain both effective and transparent.

3. Conclusion

The rise of encrypted network traffic, while vital for ensuring data privacy and security, has significantly challenged traditional intrusion detection systems by obscuring payload information. In this context, explainable artificial intelligence (XAI) has emerged as a critical enabler for advancing cybersecurity operations. By leveraging flow-level metadata and side-channel features, XAI-integrated models offer the dual advantages of robust threat detection and transparent decision-making. Techniques such as SHAP, LIME, attention mechanisms, and interpretable models like decision trees or XGBoost empower analysts to understand, verify, and act upon model outputs even in the absence of content inspection.

The inclusion of XAI in encrypted traffic analysis is especially valuable for fostering trust, auditability, and operational usability. Security analysts are more likely to act on AI-generated alerts when they can interpret the rationale behind classifications. In regulated environments, explainability ensures compliance with legal mandates for algorithmic transparency. Furthermore, interpretable systems support faster triage and incident response, reducing operational overhead and enhancing resilience against evolving threats.

Despite these advantages, the deployment of XAI in cybersecurity remains limited. There is a pressing need for broader adoption of interpretable AI techniques, not only in research prototypes but also in production-grade intrusion detection systems. Future work must focus on improving the scalability, efficiency, and usability of XAI tools, ensuring that they can be seamlessly integrated into real-time, distributed, and resource-constrained environments. As adversaries continue to evolve their tactics, adopting transparent and accountable AI systems will be essential for maintaining defensive agility and organizational trust. In sum, XAI represents not just a technical advancement, but a paradigm shift toward responsible and explainable cybersecurity in an increasingly encrypted digital landscape.

4. References

1. Abdelmoumin G, Rawat DB, Rahman A. On the performance of machine learning models for anomaly-based intelligent intrusion detection systems for the internet of things. IEEE Internet Things J. 2021;9(6):4280-90.
2. Ables J, Kirby T, Anderson W, Mittal S, Rahimi S, Banicescu I, *et al*. Creating an explainable intrusion detection system using self organizing maps. In: 2022 IEEE Symposium Series on Computational Intelligence (SSCI); 2022 Dec 4-7; Singapore. Piscataway (NJ): IEEE; 2022. p. 404-12.
3. Ahsan MM, Alam TE, Trafalis T, Huebner P. Deep MLP-CNN model using mixed-data to distinguish between COVID-19 and Non-COVID-19 patients. Symmetry. 2020;12(9):1526.
4. Aisyah N, Hidayat R, Zulaikha S, Rizki A, Yusof ZB, Pertiwi D, *et al*. Artificial intelligence in cryptographic protocols: securing e-commerce transactions and ensuring data integrity. [Place unknown]: [Publisher unknown]; 2019.
5. Akbari I, Salahuddin MA, Aniva L, Limam N, Boutaba R, Mathieu B, *et al*. A look behind the curtain: traffic classification in an increasingly encrypted web. Proc ACM Meas Anal Comput Syst. 2021;5(1):1-26.
6. Alahmadi BA, Axon L, Martinovic I. 99% false positives: a qualitative study of SOC analysts' perspectives on security alarms. In: 31st USENIX Security Symposium (USENIX Security 22); 2022 Aug 10-12; Boston (MA). Berkeley (CA): USENIX Association; 2022. p. 2783-800.
7. Alam M, Bhattacharya S, Mukhopadhyay D. Victims can be saviors: a machine learning-based detection for micro-architectural side-channel attacks. ACM J Emerg Technol Comput Syst. 2021;17(2):1-31.
8. Alkadi O, Moustafa N, Turnbull B. A review of intrusion detection and blockchain applications in the cloud: approaches, challenges and solutions. IEEE Access. 2020;8:104893-917.
9. Alonge EO, Eyo-Udo NL, Ubanadu BC, Daraojimba AI, Balogun ED, Ogunsola KO. Enhancing data security with machine learning: a study on fraud detection algorithms. J Data Secur Fraud Prev. 2021;7(2):105-18.
10. Awotunde JB, Misra S. Feature extraction and artificial intelligence-based intrusion detection model for a secure internet of things networks. In: Illumination of artificial intelligence in cybersecurity and forensics. Cham: Springer International Publishing; 2022. p. 21-44.
11. Ayodeji A, Liu YK, Chao N, Yang LQ. A new perspective towards the development of robust data-

driven intrusion detection for industrial control systems. Nucl Eng Technol. 2020;52(12):2687-98.
12. Bhatiaa A, Bahugunaa AA, Tiwaria K, Haribabua K, Vishwakarmab D. A survey on analyzing encrypted network traffic of mobile devices. arXiv. 2020; arXiv:2006.12352.
13. Bold R, Al-Khateeb H, Ersotelos N. Reducing false negatives in ransomware detection: a critical evaluation of machine learning algorithms. Appl Sci. 2022;12(24):12941.
14. Burkart P, McCourt T. Why hackers win: power and disruption in the network society. Oakland (CA): University of California Press; 2019.
15. Callegari C, Ducange P, Fazzolari M, Vecchio M. Explainable internet traffic classification. Appl Sci. 2021;11(10):4697.
16. Canavese D, Regano L, Basile C, Ciravegna G, Lioy A. Encryption-agnostic classifiers of traffic originators and their application to anomaly detection. Comput Electr Eng. 2022;97:107621.
17. Catillo M, Pecchia A, Villano U. No more DoS? An empirical study on defense techniques for web server denial of service mitigation. J Netw Comput Appl. 2022;202:103363.
18. Cherekar R. The future of data governance: ethical and legal considerations in AI-driven analytics. Int J Artif Intell Data Sci Mach Learn. 2022;3(2):17-24.
19. Choi K, Yi J, Park C, Yoon S. Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. IEEE Access. 2021;9:120043-65.
20. Dang QV. Improving the performance of the intrusion detection systems by the machine learning explainability. Int J Web Inf Syst. 2021;17(5):537-55.
21. Darias JM, Díaz-Agudo B, Recio-Garcia JA. A systematic review on model-agnostic XAI libraries. In: ICCBR workshops; 2021 Sep 13-16; [Place unknown]. [Place unknown]: [Publisher unknown]; 2021. p. 28-39.
22. Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv. 2020; arXiv:2006.11371.
23. Djenna A, Harous S, Saidouni DE. Internet of things meet internet of threats: new concern cyber security issues of critical cyber infrastructure. Appl Sci. 2021;11(10):4580.
24. Eleanor H. Modernizing data security: best practices for compliance with US and international privacy regulations. Int J Trend Sci Res Dev. 2021;5(4):1881-94.
25. Eskandari M, Janjua ZH, Vecchio M, Antonelli F. Passban IDS: an intelligent anomaly-based intrusion detection system for IoT edge devices. IEEE Internet Things J. 2020;7(8):6882-97.
26. Faith H, Agoro H. Technological solutions for CCPA compliance. [Place unknown]: [Publisher unknown]; 2022.
27. Fernandes G, Rodrigues JJ, Carvalho LF, Al-Muhtadi JF, Proença ML. A comprehensive survey on network anomaly detection. Telecomm Syst. 2019;70:447-89.
28. Fiero AW, Beier E. New global developments in data protection and privacy regulations: comparative analysis of European Union, United States, and Russian legislation. Stan J Int Law. 2022;58:151.
29. Gittens A, Yener B, Yung M. An adversarial perspective on accuracy, robustness, fairness, and privacy: multilateral-tradeoffs in trustworthy ML. IEEE Access. 2022;10:120850-65.
30. Gryz J, Rojszczak M. Black box algorithms and the rights of individuals: no easy solution to the "explainability" problem. Internet Policy Rev. 2021;10(2):1-24.
31. Gurbani V, Hood C, Nikolich A, Schulzrinne H. When DNS goes dark: understanding privacy and shaping policy of an evolving protocol. In: TPRC48: The 48th Research Conference on Communication, Information and Internet Policy; 2020 Dec; [Place unknown]. [Place unknown]: [Publisher unknown]; 2020.
32. Habeeb MS, Babu TR. Network intrusion detection system: a survey on artificial intelligence-based techniques. Expert Syst. 2022;39(9):e13066.
33. Hajj S, El Sibai R, Bou Abdo J, Demerjian J, Makhoul A, Guyeux C. Anomaly-based intrusion detection systems: the requirements, methods, measurements, and datasets. Trans Emerg Telecomm Technol. 2022;32(4):e4240.
34. Hou J, Liu H, Liu Y, Wang Y, Wan PJ, Li XY. Model protection: real-time privacy-preserving inference service for model privacy at the edge. IEEE Trans Dependable Secure Comput. 2021;19(6):4270-84.
35. Hu J, Zheng W. Multistage attention network for multivariate time series prediction. Neurocomputing. 2020;383:122-37.
36. Ibraheem HR, Zaki ND, Al-mashhadani MI. Anomaly detection in encrypted HTTPS traffic using machine learning: a comparative analysis of feature selection techniques. Mesopotamian J Comput Sci. 2022:18-28.
37. Jagannath J, Polosky N, Jagannath A, Restuccia F, Melodia T. Machine learning for wireless communications in the Internet of Things: a comprehensive survey. Ad Hoc Netw. 2019;93:101913.
38. Kummari DN. Machine learning applications in regulatory compliance monitoring for industrial operations. Glob Res Dev J. 2020;5(12):75-95.
39. Ladbury C, Zarinshenas R, Semwal H, Tam A, Vaidehi N, Rodin AS, *et al*. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. Transl Cancer Res. 2022;11(10):3853.
40. Li Y, Zhu Z, Kong D, Han H, Zhao Y. EA-LSTM: evolutionary attention-based LSTM for time series prediction. Knowl Based Syst. 2019;181:104785.
41. Liang F, Hatcher WG, Liao W, Gao W, Yu W. Machine learning for security and the internet of things: the good, the bad, and the ugly. IEEE Access. 2019;7:158126-47.
42. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy. 2020;23(1):18.
43. Lopes P, Silva E, Braga C, Oliveira T, Rosado L. XAI systems evaluation: a review of human and computer-centred methods. Appl Sci. 2022;12(19):9423.
44. Lou X, Zhang T, Jiang J, Zhang Y. A survey of microarchitectural side-channel vulnerabilities, attacks, and defenses in cryptography. ACM Comput Surv. 2021;54(6):1-37.
45. Markopoulou D, Papakonstantinou V. The regulatory framework for the protection of critical infrastructures against cyberthreats: identifying shortcomings and addressing future challenges: the case of the health sector in particular. Comput Law Secur Rev. 2021;41:105502.
46. Martins I, Resende JS, Sousa PR, Silva S, Antunes L,

Gama J. Host-based IDS: a review and open issues of an anomaly detection system in IoT. Future Gener Comput Syst. 2022;133:95-113.
47. Marwah GPK, Jain A, Malik PK, Singh M, Tanwar S, Safirescu CO, *et al*. An improved machine learning model with hybrid technique in VANET for robust communication. Mathematics. 2022;10(21):4030.
48. Meryem A, Ouahidi BE. Hybrid intrusion detection system using machine learning. Netw Secur. 2020;2020(5):8-19.
49. Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev. 2022:1-66.
50. Mintoo AA, Saimon ASM, Bakhsh MM, Akter M. National resilience through AI-driven data analytics and cybersecurity for real-time crisis response and infrastructure protection. Am J Sch Res Innov. 2022;1(01):137-69.
51. Kacheru G. The future of cyber defence: predictive security with artificial intelligence. Int J Adv Res Basic Eng Sci Technol. 2021;7(12):46-55.
52. Moreno V, Génova G, Parra E, Fraga A. Application of machine learning techniques to the flexible assessment and improvement of requirements quality. Softw Qual J. 2020;28(4):1645-74.
53. Mousavi SK, Ghaffari A, Besharat S, Afshari H. Security of internet of things based on cryptographic algorithms: a survey. Wirel Netw. 2021;27(2):1515-55.
54. Nasir V, Sassani F. A review on deep learning in machining and tool monitoring: methods, opportunities, and challenges. Int J Adv Manuf Technol. 2021;115(9):2683-709.
55. Neupane S, Ables J, Anderson W, Mittal S, Rahimi S, Banicescu I, *et al*. Explainable intrusion detection systems (X-IDS): a survey of current methods, challenges, and opportunities. IEEE Access. 2022;10:112392-415.
56. Ogrezeanu I, Vizitiu A, Ciușdel C, Puiu A, Coman S, Boldișor C, *et al*. Privacy-preserving and explainable AI in industrial applications. Appl Sci. 2022;12(13):6395.
57. Oye E, Mitchell N. Future of data privacy regulations beyond CCPA. [Place unknown]: [Publisher unknown]; 2022.
58. Papadogiannaki E, Ioannidis S. A survey on encrypted network traffic analysis applications, techniques, and countermeasures. ACM Comput Surv. 2021;54(6):1-35.
59. Papadogiannaki E, Tsirantonakis G, Ioannidis S. Network intrusion detection in encrypted traffic. In: 2022 IEEE Conference on Dependable and Secure Computing (DSC); 2022 Jun 22-24; [Place unknown]. Piscataway (NJ): IEEE; 2022. p. 1-8.
60. Ramlakhan S, Saatchi R, Sabir L, Singh Y, Hughes R, Shobayo O, *et al*. Understanding and interpreting artificial intelligence, machine learning and deep learning in emergency medicine. Emerg Med J. 2022;39(5):380-5.
61. Rehan H. Leveraging AI and cloud computing for real-time fraud detection in financial systems. J Sci Technol. 2021;2(5):127.
62. Rimmer V, Nadeem A, Verwer S, Preuveneers D, Joosen W. Open-world network intrusion detection. In: Security and artificial intelligence: a crossdisciplinary approach. Cham: Springer International Publishing; 2022. p. 254-83.
63. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.
64. Russo M, Šrndić N, Laskov P. Detection of illicit cryptomining using network metadata. EURASIP J Inf Secur. 2021;2021(1):11.
65. Sahoo BB, Jha R, Singh A, Kumar D. Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophys. 2019;67(5):1471-81.
66. Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. Towards a standard for identifying and managing bias in artificial intelligence. Gaithersburg (MD): US Department of Commerce, National Institute of Standards and Technology; 2022. Report No.: NIST SP 1270.
67. Shah H. Artificial intelligence with safe and secure deep learning architectures. Int Res J Eng Appl Sci. 2019;7(3):10-55083.
68. Lewechi FE. Zero trust framework for AI-enabled digital twin: integrating security, fairness, and compliance monitoring. Int J Multidiscip Res Growth Eval. 2023;4(6):1339-1347. doi:10.54660/IJMRGE.2023.4.6.1339-1347.
69. Lewechi F. Blockchain-orchestrated IAM for multi-cloud AI systems: identify federation with ethical controls. Int J Multidiscip Evolut Res. 2023;4(2):139–149. doi:10.54660/IJMER.2023.4.2.139-149.
70. Shankar R, Ahmed F. Explainable AI (XAI): methods, tools, and challenges in interpreting machine learning models. Artif Intell Mach Learn Rev. 2021;2(1):1-9.
71. Sheu RK, Pardeshi MS. A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. Sensors. 2022;22(20):8068.
72. Sun J, Sun K, Shenefiel C. Automated IoT device fingerprinting through encrypted stream classification. In: International Conference on Security and Privacy in Communication Systems; 2019 Oct; [Place unknown]. Cham: Springer International Publishing; 2019. p. 147-67.
73. Townsend J, Chaton T, Monteiro JM. Extracting relational explanations from deep neural networks: a survey from a neural-symbolic perspective. IEEE Trans Neural Netw Learn Syst. 2019;31(9):3456-70.
74. Wickramasinghe CS, Amarasinghe K, Marino DL, Rieger C, Manic M. Explainable unsupervised machine learning for cyber-physical systems. IEEE Access. 2021;9:131824-43.
75. Wood P. A global perspective. In: Routledge handbook of risk management and the law. Abingdon: Routledge; 2022. p. 14.
76. Yasasin E, Prester J, Wagner G, Schryen G. Forecasting IT security vulnerabilities – an empirical analysis. Comput Secur. 2020;88:101610.
77. Zakhary S, Lodge T, McAuley D. Performance evaluation for privacy-preserving control of domestic IoT devices. arXiv. 2022; arXiv:2207.08482.
78. Zave P, Rexford J. Patterns and interactions in network security. ACM Comput Surv. 2020;53(6):1-37.
79. Zhang C, Patras P, Haddadi H. Deep learning in mobile and wireless networking: a survey. IEEE Commun Surv Tutor. 2019;21(3):2224-87.